(4)

The Pennsylvania State University

Department of Statistics

University Park, Pennsylvania

TECHNICAL REPORTS AND PREPRINTS

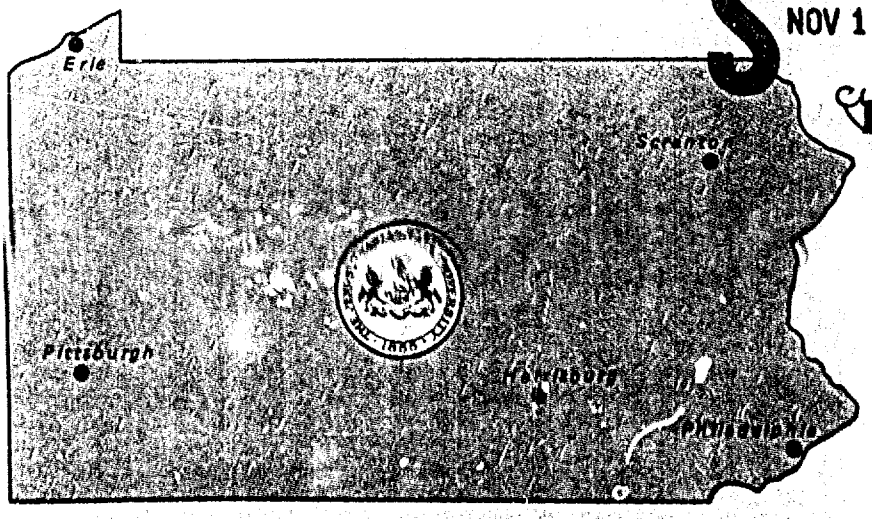
Number 77: November 1988

ROBUST BOUNDED INFLUENCE TESTS IN LINEAR MODELS

Marianthi Markatou
The University of Iowa

and

Thomas P. Hettmansperger*
The Pennsylvania State University



(4)

DEPARTMENT OF STATISTICS

The Pennsylvania State University
University Park, PA 16802 U.S.A.

TECHNICAL REPORTS AND PREPRINTS

Number 77: November 1988

ROBUST BOUNDED INFLUENCE TESTS IN LINEAR MODELS

Marianthi Markatou
The University of Iowa

and

Thomas P. Hettmansperger*
The Pennsylvania State University

*Research partially supported by ONR contract N00014-80-C0741.

Title: Robust Bounded influence tests in linear models

Authors: Marianthi Markatou
The University of Iowa, Iowa City, Iowa

Thomas P. Hettmansperger*
The Pennsylvania State University, University Park

1. SUMMARY

A robust test, which we call an aligned generalized, M–test for testing subhypotheses in the general linear models is developed, and its asymptotic properties are studied. The test is a robustification of the well known F–test, and it is an elegant and practical alternative to Ronchetti's (1982) class of τ–tests. P–values associated with it can be approximated readily using existing chi square tables, unlike Ronchetti's test. The test is based on an appropriately constructed quadratic form, and uses the generalized M–estimators of the parameters in the reduced model. Under the null hypothesis the asymptotic distribution is a central chi square, and under contiguous alternatives is a non–central chi square with the same degrees of freedom. The test can also be viewed as a generalization of Sen's (1982) M–test for linear models.

*Research partially supported by ONR contract N00014–80–CO741.

The influence function of the test is bounded. The bound not only applies to the influence of residuals but also to the influence of position in the factor space. Sen's test, on the other hand, has bounded influence only in residuals. (KR)

Some key words: aligned generalized M–test; bounded influence; contiguous alternatives; influence function; linear model; random carriers; robustness; τ–test.

2. INTRODUCTION AND THE MODEL

We consider the following model: $\{(\underline{x}_i, y_i);\ i = 1,2,\cdots,n\}$ are independent random variables such that

$$y_i = \underline{x}_i^T \underline{\beta} + \epsilon_i \tag{2.1}$$

where $\underline{x}_i$ is independent of ϵ_i and has distribution function $K(\underline{x})$ with density $k(\underline{x})$. In the above linear model it is well known that least squares estimates are very sensitive to aberrant data points, that is points that deviate significantly from the bulk of the data. This sensitivity has led to various proposals for robust methods of estimation. Among those proposals are the classical M–estimates introduced by Huber (1973), and the generalized M–estimates introduced by Hampel (1977) and discussed by Maronna and Yohai (1981).

Parameter estimation is the first step in data analysis. Often, we are interested in testing if a number of linearly independent estimable functions are equal to zero. Through a transformation in the parameter space this hypothesis

reduces to the hypothesis of testing if a certain subvector of the vector of unknown parameters equals zero, treating the remaining parameters as nuisance parameters. Known robust testing procedures are the p_c–test, introduced by Schrader and Hettmansperger (1980), which makes use of Huber's M–estimators, and the alternative to the p_c–test, an aligned M–test, which was introduced by Sen (1982).

Let $\beta^T = (\beta_1^T \beta_2^T)$ where β is a $p \times 1$ vector and β_2 is a $q \times 1$ vector, $q < p$. The above testing procedures, for the hypothesis $H_0 : \beta_2 = \underline{0}$, β_1 unspecified, are robust against points that exhibit large residuals. The influence with respect to the residuals of the above testing procedures is bounded. Their total influence though is unbounded, since the part of it that corresponds to the influence in the factor space is unbounded (see Hampel et al., 1986, page 354).

Ronchetti (1982) introduced the class of τ–tests which makes use of the generalized M–estimates, which have bounded influence. The class of τ–tests, for testing the hypothesis $H_0 : \beta_2 = \underline{0}$, β_1 unspecified, is defined by means of the test statistic

$$S_n^2(\underline{x}_1, \cdots, \underline{x}_n; y_1, \cdots, y_n) := \tag{2.2}$$

$$= 2q^{-1}n^{-1} \sum_{i=1}^{n} \left\{ \tau\left[\underline{x}_i; \frac{y_i - \underline{x}_i^T (\underline{T}_\omega)_n}{\sigma}\right] - \tau\left[\underline{x}_i; \frac{y_i - \underline{x}_i^T (\underline{T}_\Omega)_n}{\sigma}\right] \right\}$$

where $(\underline{T}_\omega)_n$, $(\underline{T}_\Omega)_n$ are the generalized M–estimates in the reduced and full model respectively, $\tau : \mathbb{R}^p \times \mathbb{R} \longrightarrow \mathbb{R}^+$ is a function such that (i) for all $\underline{x} \in \mathbb{R}^p$, $r \in \mathbb{R}$, $\tau(\underline{x};r) \not\equiv 0$, $\tau(\underline{x};r) \geq 0$, $\tau(\underline{x};0) = 0$, and (ii) for all $\underline{x} \in \mathbb{R}^p$, $\tau(\underline{x};\cdot)$ is differentiable. See Hampel et al. (1986, p. 345) for a complete set of regularity conditions. Large values of S_n^2 are significant.

The asymptotic distribution of Ronchetti's (1982) τ–test statistic, under H_0, is given by the distribution of

$$L = \sum_{j=p-q+1}^{p} \lambda_j N_j^2 \tag{2.3}$$

where $\lambda_{p-q+1} \geq \lambda_{p-q+2} \geq \cdots \geq \lambda_p > 0$ are the q positive eigenvalues of the matrix

$$K = Q\left[M^{-1} - \begin{bmatrix} M_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right] \tag{2.4}$$

where M_{11}^{-1} is the inverse of the upper $(p-q) \times (p-q)$ part of the M matrix, and

$$M = E\{\eta'(\underline{x};r)\underline{x}\,\underline{x}^T\}, \quad Q = E\{\eta^2(\underline{x};r)\underline{x}\,\underline{x}^T\} \tag{2.5}$$

where

$$\frac{\partial}{\partial r}\tau(\underline{x};r) = \eta(\underline{x};r) \quad , \quad \frac{\partial}{\partial r}\eta(\underline{x};r) = \eta'(\underline{x};r). \tag{2.6}$$

Further, N_j, $j = p-q+1, \cdots, p$ are independent standard normal random variables. The most important choices of the η–function are of the form

$$\eta(\underline{x};\frac{r}{\sigma}) = v(\underline{x})\psi_c\left[\frac{r}{\sigma v(\underline{x})}\right] \tag{2.7}$$

where $\nu : \mathbb{R}^p \longrightarrow \mathbb{R}^+$ is a weight function, $\sigma > 0$ and r is the residual, and $\psi_c(t) = \max[-c,\min(c,t)]$, the Huber ψ–function.

The τ–test can be viewed as a complete robustification of the well known F–test, since it accommodates points with large residuals and points with high leverage. It is impractical however, since it is difficult to calculate the p–values associated with it.

We will develop an alternative test, which can be viewed as a practical alternative to the class of τ–tests. We call it an aligned generalized M–test, and it is based on a properly constructed quadratic form which makes use of the generalized M–estimates. The bounded influence property of the generalized M–estimators carries over to the test based on them. Thus the influence function of the aligned generalized M–test is bounded.

The asymptotic distribution of the aligned generalized M–test, under the null hypothesis, is that of a central chi square random variable with q degrees of freedom. Therefore p–values associated with the test statistic can be approximated readily from existing chi square tables. The asymptotic distribution of the test, under contiguous alternatives, is that of a noncentral chi square random variable.

3. THE PROPOSED TEST, ASSUMPTIONS AND DISTRIBUTION THEORY

The model is that of section 2. We will assume the regularity conditions (A1) – (A6) and (C1) – (C6) listed in Maronna and Yohai (1981). The notation is slightly different, however we will not reproduce all the conditions here.
In addition, we assume that the density of the residuals is bounded. Without loss of

generality, assume that $\sigma = 1$. In practice σ has to be estimated from the data. We will discuss the estimation of the scale parameter σ in a latter section.

Estimators based on (2.7) have been studied by Hampel (1977) and Krasker (1980) for the special case of $\nu(\underline{x}) = 1/||\underline{x}||$, where $||\cdot||$ is the Euclidean norm of $\underline{x}$, and for the general case by Maronna and Yohai (1981).

We will use generalized M–estimators, defined implicitly by the (vector) equation

$$\sum_{i=1}^{n} \eta\left[\underline{x}_i;\, y_i - \underline{x}_i^T \underline{\beta}\right]\underline{x}_i = \underline{0} \tag{3.1}$$

Define the dispersion function

$$D(\underline{\beta}) = \sum_{i=1}^{n} \tau(\underline{x}_i;\, y_i - \underline{x}_i^T \underline{\beta})\underline{x}_i. \tag{3.2}$$

Its gradient, with respect to $\underline{\beta}$, is

$$\nabla D(\underline{\beta}) = -\sum_{i=1}^{n} \eta(\underline{x}_i;\, y_i - \underline{x}_i^T \underline{\beta})\underline{x}_i. \tag{3.3}$$

Let

$$\underline{S}(\beta) = -\nabla D(\beta) = \sum_{i=1}^{n} \eta(\underline{x}_i;\, y_i - \underline{x}_i^T \beta)\underline{x}_i. \tag{3.4}$$

Let $\underline{x}_i^T = (\underline{x}_i^{*T}\ \underline{x}_i^{**T})$, where $\underline{x}_i^{**}$ is a q–dimensional vector, be the corresponding partition to the partition of the vector of unknown parameters. To test the hypothesis $H_0 : \beta_2 = \underline{0},\ \beta_1$ unspecified versus $H_1 : \beta_2 \neq \underline{0},\ \beta_1$ unspecified, define the statistic W_n^2 as follows:

$$W_n^2 = \left[n^{-1/2} \sum_{i=1}^{n} \eta(\underline{x}_i; y_i - \underline{x}_i^T \hat{\beta}_0)\underline{x}_i^{**}\right]^T \hat{U}^{-1} \left[n^{-1/2} \sum_{i=1}^{n} \eta(\underline{x}_i; y_i - \underline{x}_i^T \hat{\beta}_0)\underline{x}_i^{**}\right] \tag{3.5}$$

where $\hat{\beta}_0$ is the reduced model generalized M–estimator, $\hat{U}$ is a consistent estimate of the asymptotic variance–covariance matrix, which is defined in Theorem 1.

Large values of W_n^2 are significant. The test statistic W_n^2 defines an aligned test, which we call an aligned generalized M–test. We will discuss the distribution theory of the test statistic W_n^2. Define

$$\underline{L}(\beta) = \underline{S}(\beta_0) - n\hat{M}(\beta - \beta_0) \tag{3.6}$$

where β_0 is the true parameter and

$$\hat{M} = \frac{1}{n} \sum_{i=1}^{n} \eta'(\underline{x}_i; y_i - \underline{x}_i^T \beta_0)\underline{x}_i \underline{x}_i^T \tag{3.7}$$

is a consistent estimator of the matrix M defined in (2.5), and η' is defined in (2.6). Assuming the regularity conditions of Maronna and Yohai (1981), the following results hold. Proofs are given in Markatou (1988).

Lemma 1

Given $K > 0$

$$\sup_{\sqrt{n}||\underline{\beta}-\underline{\beta}_0||\leq K} ||n^{-1/2}\underline{S}(\underline{\beta}) - n^{-1/2}\underline{L}(\underline{\beta})|| \tag{3.8}$$

converges in probability to zero.

Lemma 2

Consider the random variables $\eta(\underline{x}_i; y_i - \underline{x}_i^T\underline{\beta})x_{ij}$ $i = 1,\cdots,n$ and $j = 1,2,\cdots,p$ where the $(y_i,\underline{x}_i)$ are independent and identically distributed random vectors. Then, under the assumptions of finite variance and zero mean

$$n^{-1/2}\underline{S}(\underline{\beta}_0) = n^{-1/2}\sum_{i=1}^{n} \eta(\underline{x}_i;\epsilon_i)\underline{x}_i$$

converges in distribution to a random variable $\underline{Z}$, where $\underline{Z} \sim MVN(\underline{0},Q)$ with Q defined in (2.5).

Lemma 3

$n^{1/2}(\hat{\underline{\beta}} - \underline{\beta}_0)$ converges in distribution to the random variable $\underline{Z}$ where

$\underline{Z} \sim MVN(\underline{0}, M^{-1}QM^{-1})$, with M and Q given in (2.5) and $\hat{\beta}$ is the generalized M–estimator in the full model.

Theorem 1

Under the null hypothesis

$$H_0 : \beta = \beta_0 = \begin{bmatrix} \beta_{01} \\ \underline{0} \end{bmatrix}$$

the test statistic $W_n^2 = n^{-1/2}\underline{S}_2^T(\hat{\beta}_0)\hat{U}^{-1}n^{-1/2}\underline{S}_2(\hat{\beta}_0)$ has asymptotically a chi square distribution with q degrees of freedom, where

$$\hat{U} = \hat{Q}_{22} - \hat{M}_{21}\hat{M}_{11}^{-1}\hat{Q}_{12} - \hat{Q}_{21}\hat{M}_{11}^{-1}\hat{M}_{12} + \hat{M}_{21}\hat{M}_{11}^{-1}\hat{Q}_{11}\hat{M}_{11}^{-1}\hat{M}_{12} \qquad (3.9)$$

and

$$\underline{S}_2(\hat{\beta}_0) = \sum_{i=1}^{n} \eta(\underline{x}_i; y_i - \underline{x}_i^T\hat{\beta}_0)\underline{x}_i^{**} \qquad (3.10)$$

and $\hat{\beta}_0$ is the generalized M–estimator in the reduced model.

3.1 DISTRIBUTION UNDER CONTIGUOUS ALTERNATIVES

We use contiguity as it is defined by Hajek and Sidak (1967, page 202). We regard the given testing problem as a member of a sequence $\{H_{0n}, H_{1n}\}$, $n \geq 1$ of testing problems, and we consider alternatives

$$H_{1n} : \beta_j = n^{-1/2}\Delta_j, \quad j = p-q+1, \cdots, p,$$

where $\underline{\Delta}^T = (\Delta_{p-q+1}, \cdots, \Delta_p)$ is a $q \times 1$ vector of finite constants. Then:

$$H_{0n} : \underline{\beta}_n = \begin{bmatrix} \underline{\beta}_{1n} \\ \underline{0} \end{bmatrix} \quad \text{vs}$$

$$H_{1n} : \underline{\beta}_n = \begin{bmatrix} \underline{\beta}_{1n} \\ n^{-1/2}\,\underline{\Delta} \end{bmatrix} = \begin{bmatrix} \underline{\beta}_{1n} \\ \underline{0} \end{bmatrix} + n^{-1/2} \begin{bmatrix} \underline{0} \\ \underline{\Delta} \end{bmatrix}$$

$$= \begin{bmatrix} \underline{\beta}_{1n} \\ \underline{0} \end{bmatrix} + \underline{\delta}_n$$

Theorem 2

Under the contiguous alternatives

$$H_{1n} : \underline{\beta}_n = \begin{bmatrix} \underline{\beta}_{1n} \\ \underline{0} \end{bmatrix} + n^{-1/2} \begin{bmatrix} \underline{0} \\ \underline{\Delta} \end{bmatrix}$$

the test statistics $n^{-1/2}\underline{S}_2^T(\hat{\underline{\beta}}_0)\hat{U}^{-1}n^{-1/2}\underline{S}_2(\hat{\underline{\beta}}_0)$ has asymptotically a noncentral

chi square distribution with q degrees of freedom, where δ is the noncentrality parameter defined by

$$\delta = \Delta^T M_{22.1} U^{-1} M_{22.1} \Delta \tag{3.11}$$

and

$$M_{22.1} = M_{22} - M_{21} M_{11}^{-1} M_{12}. \tag{3.12}$$

Sen (1982), introduced the aligned M–test in linear models, based on the classical M–estimators. Note that, if $\eta(\underline{x};r) = \psi_c(r)$ where $\psi_c(r) = \max\{-c,\min(c,r)\}$ is Huber's phi–function, and r denotes a residual, the above aligned generalized M–test reduces to Sen's M–test. Therefore, Sen's M–test is a special case of the aligned generalized M–test. The noncentrality parameter δ, in the case of simple linear regression and with $\eta(\underline{x};r) = \psi_c(r)$, is given as

$$\delta = \frac{\lambda^2 \gamma^2 \sigma_x^2}{\sigma_0^2} \tag{3.13}$$

where $\lambda = \Delta$, $\gamma = E[\psi_c'(r)]$ and $n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \xrightarrow{p} \sigma_x^2$ that is Sen's noncentrality parameter.

4. INFLUENCE FUNCTION

We calculate the influence function of the aligned generalized M–test and show that it is bounded. Roughly speaking, the influence function formalizes the bias caused by one outlier. It was introduced by Hampel (1968, 1974) and it describes the effect of an infinitesimal contamination at a point on the estimate, standardized by the mass of the contamination.

Its formal definition is as follows:

$$IF(x;T,F) = \lim_{t \downarrow 0} \frac{T((1-t)F+t\Delta_x)}{t}$$

for those $x \in \mathcal{X}$ where the limit exists. T is an estimator which is a functional or asymptotically can be replaced by a functional, F is a distribution function and Δ_x is the distribution that concentrates all the mass at x.

The influence function of a test statistic describes the effect of an outlier in the sample on the value of the (standardized) test statistic, and therefore on the decision (acceptance or rejection of H_0) which is based on this value. It is analogous to the influence function of the estimators, which the test is based upon.

The influence function of the aligned generalized M–test is calculated under the null hypothesis $H_0 : \beta_2 = \underline{0}$, β_1 unspecified. It is proved that, under certain conditions on the η–function, the influence function is bounded. Let $F_{\underline{\beta}}(\underline{x};y)$ denote the joint distribution function of $(\underline{x},y)$.

Note that the model distribution under H_0 is F_{β_0}. Define:

$$J(\beta) = \int \eta(x;y-x^T\beta)x^{**}dF_\beta(x;y) \tag{4.1}$$

Theorem 3

Let W_n be the aligned generalized M–test, and assume the regularity conditions of Maronna and Yohai (1981) are satisfied. Assume $J(\beta_0) = 0$ (Fisher consistency). Then, the influence function of the test statistic at (x_0,y_0) is given as:

$$IF(x_0,y_0;W,F_{\beta_0}) = |\eta(x_0;y_0-x_0^T\beta_0)| \left[x_0^T \begin{bmatrix} -M_{11}^{-1}M_{12} \\ I \end{bmatrix} U^{-1}(-M_{21}M_{11}^{-1} \;\; I)x_0\right]^{1/2}.$$

The proof of Theorem 3 is similar to the proof of Proposition 1 in Hampel et al. (1986, p. 350) and hence we do not repeat it here; see also Markatou (1988).

The most important choices for the function $\eta(x;r)$ are of the form given by (2.7). In order for the influence function to be bounded $v(x_0)x_0^TAx_0$ must be bounded, where

$$A = \begin{bmatrix} -M_{11}^{-1}M_{12} \\ I \end{bmatrix} U^{-1}(-M_{21}M_{11}^{-1} \;\; I). \tag{4.2}$$

Note that the matrix A does not depend on the outlying case (x_0,y_0). Thus a natural choice for the η–function would be the one for which

$$\nu(\underset{\sim}{x}_0) = \left[\underset{\sim}{x}_0^T \begin{bmatrix} -M_{11}^{-1}M_{12} \\ I \end{bmatrix} U^{-1}(-M_{21}M_{11}^{-1} \; I)\underset{\sim}{x}_0\right]^{-1/2}. \tag{4.3}$$

In that case

$$\sup_{\underset{\sim}{x},y} |IF(\underset{\sim}{x},y;W,F_{\underset{\sim}{\beta}_0})| \leq c < \infty \tag{4.4}$$

where c is a finite constant. See, for example, Krasker and Welsch (1982).

There are other choices of $\nu(\underset{\sim}{x})$ for which the influence function of the test statistic remains bounded.

(1) Following Schweppe's proposal introduced in Handshin et al. (1975), we choose $\nu(\underset{\sim}{x}_0) = (1-\underset{\sim}{x}_0^T(X^TX)^{-1}\underset{\sim}{x}_0)^{1/2} = (1-h_0)^{1/2}$. Then, as $h_0 \longrightarrow 1$, $\nu(x_0) \longrightarrow 0$, where h_0 is the leverage that corresponds to the outlying case $(\underset{\sim}{x}_0,y_0)$. Then, the sensitivity of the test statistic W is

$$\sup_{\underset{\sim}{x},y}|IF(\underset{\sim}{x},y;W,F_{\underset{\sim}{\beta}_0})| \leq \sup_{\underset{\sim}{x},y}||(-M_{21} \; M_{22})IF(\underset{\sim}{x},y;\underset{\sim}{\beta}_0,F_{\underset{\sim}{\beta}_0})||$$

$$\leq c_1 < \infty$$

that is, the sensitivity of the test is bounded by a rescaled version of the sensitivity of the reduced model generalized M–estimator.

(2) Welsch's (1980) proposal consists of choosing $\nu(\underset{\sim}{x}_0) = [1 - \underset{\sim}{x}_0^T(X^TX)^{-1}\underset{\sim}{x}_0]/[\underset{\sim}{x}_0^T(X^TX)^{-1}\underset{\sim}{x}_0]^{1/2} = (1-h_0)/h_0^{1/2}$. Then, as $h_0 \longrightarrow 1$, $\nu(\underset{\sim}{x}_0) \longrightarrow 0$ and thus, the influence function of the test statistic reduces to a rescaled

version of the influence function of the reduced model generalized M–estimate, which has bounded sensitivity.

Generally, starting with a bounded influence estimator, the influence function of the test based on this estimator will be bounded.

5. EXAMPLE

The data design comes from a paper by Hill and Holland (1977) and consists of six columns. Table 5.1 contains the values of the explanatory variables in its first six columns. Column number 7 corresponds to the data values of the dependent variable y.

Table 5.1: Data Values

Column \Row	1	2	3	4	5	6	7
1	.2712	.2712	−.0453	.0257	−.0880	.0288	.92718
2	.2712	.1627	.1092	−.1268	−.0509	.0470	−.06165
3	.2712	.0542	.4513	.0963	.0140	.0682	−.74198
4	.2712	−.0542	−.1605	.2977	−.1065	.0225	−.31325
5	.2712	−.1627	.2242	−.3618	.2463	.3193	−.18593
6	.2712	−.2712	.0107	.1246	−.0814	.0461	.18593
7	.1627	−.2712	.1937	.1006	−.0373	.0583	.31325
8	.0542	−.2712	−.2435	.3205	−.1373	.0404	−1.40377
9	−.0542	−.2712	−.0094	−.4123	−.0852	.0228	1.12690
10	−.1627	−.2712	.1382	.4631	−.0630	−.0112	1.87129
11	−.2712	−.2712	.0956	.0984	−.0489	.0388	−.91718
12	−.2712	−.1627	.0597	−.1136	−.0732	.0327	.74198
13	−.2712	−.0542	−.0613	−.1263	−.0944	.0303	.06165
14	−.2712	.0542	.1282	.0598	−.0680	.0691	−1.87129
15	−.2712	.1627	−.0966	−.0085	.1387	−.0672	−.58740
16	−.2712	.2712	−.1060	−.3819	−.1340	.0559	−.44602
17	−.1627	.2712	.2013	.0145	−.0290	.0966	1.40377
18	.0542	.2712	.0914	.0840	−.0417	−.0620	.58740
19	−.0542	.2712	−.4324	−.2083	−.1520	−.9198	−1.12690
20	.1627	.2712	−.5486	.0544	.8917	.0833	.44602

The model to be fitted is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i \qquad (5.1)$$

$$i = 1, \cdots, 20.$$

In vector formulation the above model can be written as:

$$\underset{\sim}{Y} = X \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{\beta}_2 \end{bmatrix} + \underset{\sim}{\epsilon} \qquad (5.2)$$

where $\underset{\sim}{\beta}_1^T = (\beta_0, \beta_1, \beta_2)$ and $\underset{\sim}{\beta}_2^T = (\beta_3, \beta_4, \beta_5, \beta_6)$. We are interested in testing the hypothesis

$$H_0 : \underset{\sim}{\beta}_2 = \underset{\sim}{0}, \ \underset{\sim}{\beta}_1 \text{ unspecified vs}$$
$$H_1 : \underset{\sim}{\beta}_2 \neq \underset{\sim}{0}, \ \underset{\sim}{\beta}_1 \text{ unspecified.}$$

In an analysis of variance setting we can interpret the above hypothesis as testing for the significance of the covariates in a designed experiment.

The reason for choosing this particular data set is that the observations exhibit varying degrees of leverage. The first two columns of the data set correspond to variables like those in a designed experiment, and hence they represent a very well behaved low leverage situation. The next two columns were selected to represent a sample from a bivariate normal distribution, and they are also low leverage observations.

Columns 5 and 6 were chosen to represent a sample of 20 observations from a

distribution with outliers. Two independent Cauchy samples of size 20 were drawn and the largest observations in each sample were moved until they contributed 80% and 85% to the total sum of squares of their columns, respectively. Thus, those two columns represent a high leverage situation. The two high leverage cases are located in rows 19 and 20.

Note that the leverage of the nineteenth case is $h_{19} = .952986$ and that of the twentieth case is $h_{20} = .747831$. If $h_i \geq 2(p/n)$, where (p/n) is the average leverage, p is the number of parameters and n is the sample size, then we characterize the i^{th} observation as a high leverage observation; see Belsley et al. (1980). In this particular example $2(p/n) = .7$, and so clearly observations 19 and 20 are high leverage ones.

After the above six columns were selected, each column was standardized to have mean 0 and unit sum of squares.

To generate a set of dependent variables $\underline{y}$ (column 7 in table 5.1), we generated 20 normal scores. Regressing the obtained set of normal scores onto the six columns generated before, we selected the random permutation of the normal scores that gives us a small R^2–coefficient. Therefore the experiment has been constructed so that no regression effect is present.

We would like to show how a combination of high leverage cases with outliers in the dependent variables distorts the results of an analysis based, not only on the classical F–test procedure, but also on robust procedures that do not accommodate high leverage cases. To this end we replace the dependent value that corresponds to the nineteenth case by $y_{19} = -7.84$. This new value is approximately six standard deviations away from the original value. The least

squares regression with $y_{19} = -7.84$ gives $R^2 = 72.6\%$ and an R^2–adj = 60%.

Table 5.2 contains the value of the various statistics and their p–values:

Table 5.2: Statistics and their p–values

Statistic	Numerical Value	P–value
F–statistic	7.9307475	.0008
M–test	13.86	.008
aligned generalized M–test	3.69264	.4492

To calculate the M–test we used Huber's ψ–function with $c = 1.345$ (that gives 95% efficiency at the normal model). Hence, in this example, the bounded influence aligned generalized M–test out performed both the F–test and the M–test.

To calculate the aligned generalized M–test we chose to use Welsch's weights; that is, the weight function $\nu(\underline{x})$ is given by

$$\nu(\underline{x}) = \frac{1-\underline{x}^T(X^TX)^{-1}\underline{x}}{(\underline{x}^T(X^TX)^{-1}\underline{x})^{1/2}} = \frac{1-h}{h^{1/2}} \tag{5.3}$$

where $h = \underline{x}^T(X^TX)^{-1}\underline{x}$. Then

$$\eta(\underline{x}_i; \frac{r_i}{s(i)}) = \begin{cases} \frac{r_i}{s(i)} & \text{if } |d_i| \le c \\ c\nu(x_i)\,\text{sign}\left[\frac{r_i}{s(i)}\right] & \text{if } |d_i| > c \end{cases} \tag{5.4}$$

where s(i) is the standard deviation calculated without the i^{th} case and is given as

$$(n-p-1)s^2(i) = (n-p)s^2 - \frac{r_i^2}{1-h_i} \tag{5.5}$$

h_i is the leverage of the i^{th} case, r_i is the residual of the i^{th} case and d_i is the corresponding DFFITS, which is defined as

$$d_i = \text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_i(i)}{h_i^{1/2} s(i)} = \left[\frac{h_i}{1-h_i}\right]^{1/2} t_i \tag{5.6}$$

t_i is the externally studentized residual (see Belsley, Kuh and Welsch, 1980). The DFFITS can be interpreted as the change in the fitted value that results when the i^{th} case is deleted and the difference is scaled by an estimate of the standard deviation of the fitted value.

The calculations are carried out using iterated least squares by making the identification, using (2.7),

$$\begin{aligned} \eta(\underline{x}; \frac{r}{\sigma}) &= w(\underline{x}, r)\frac{r}{\sigma} \\ &= \min\left\{1, \frac{c\sigma v(\underline{x})}{|r|}\right\} \frac{r}{\sigma} \end{aligned} \tag{5.7}$$

Then

$$M = E\left\{\frac{1}{\sigma}\,\psi_c'\left[\frac{r}{\sigma v(\underline{x})}\right]\underline{x}\,\underline{x}^T\right\} \tag{5.8}$$

where $\psi_c'(t) = I(|t| \le c)$ and

$$Q = E\{w^2(\underline{x};r)(\frac{r}{\sigma})^2\underline{x}\,\underline{x}^T\} \tag{5.9}$$

Estimates are:

$$\hat{M} = \frac{1}{n\hat{\sigma}}\sum_{i=1}^{n}\psi_c'\left[\frac{r_i}{\hat{\sigma}v(\underline{x}_i)}\right]\underline{x}_i\,\underline{x}_i^T \tag{5.10}$$

and

$$= \frac{1}{n\hat{\sigma}^2}\sum_{i=1}^{n} w^2(\underline{x}_i;r_i)r_i^2\underline{x}_i\underline{x}_i^T \tag{5.11}$$

In the weighted least squares formulation using Welsch's weights, cases are smoothly downweighted according to how much $|d_i|$ exceeds c. Following the recommendation in Belsley, Kuh, and Welsch (1980), we took $c = (p/n)^{1/2} = 1.183$. The value of the aligned generalized M–test statistic is equal to 3.69 with an approximate p–value of .4492. Thus we fail to reject the null hypothesis. On the other hand, the F and M tests strongly reject H_0 and yield misleading results.

To estimate the scale parameter σ for the calculation of the M–test we can either use Huber's proposal 2 (Huber, 1981) or the estimator $\hat{\sigma} = 2.1 * \text{med}\{|r_i^*|\}$, where $|r_i^*|$ are the n–p+1 largest absolute values of the residuals (Hill and

Holland, 1977). In our calculation of the M–test we used the Hill and Holland estimator of the scale.

6. SUMMARY AND CONCLUSIONS

The effect that the presence of a combination of highly influential points, that is points of high leverage and outliers in the y variable, has on the testing procedures has been studied.

In the least squares context, aberrant cases can determine the estimators of the true parameter vector. They confuse the results of the testing procedure based on the least squares estimators, since the test reflects the contribution of those individual points in the model.

The M–test, in this context, is unreliable, as well as the F–test, though the M–test shows a better behavior than the F–test with a p–value .008 compared to .0008 of the F–test. The better behavior of the M–test can be explained by the fact that the test accommodates large residual points; but it does not accommodate high leverage points.

The aligned generalized M–test shows the best behavior. Designed to accommodate points with large residuals as well as high leverage points, it does give reliable results in the presence of the above mentioned combination. Its influence function is bounded. It agrees also with the τ–test, to which it is a practical alternative, since one does not face the problem of computing p–values associated with linear combinations of differentially weighted chi square random variables.

Thus, an elegant, computationally easy and reliable alternative to the class of τ–tests is the class of aligned generalized M–tests.

REFERENCES

1. Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Source of Collinearity. Wiley, New York.
2. Hajek, J. and Sidak, Z. (1967). Theory of rank tests. Academic Press, New York.
3. Hampel, F.R. (1968). Contributions to the theory of robust estimation. Ph.D Thesis, University of California, Berkeley.
4. Hampel, F.R. (1977). Modern trends in the theory of robustness. Research Report 13, ETH Zürich.
5. Hampel, F.R. (1978). Optimally bounding the gross error sensitivity and the influence of position in factor space. Proceedings of the ASA Statistical Computing Section, ASA, Washington, D.C., pp. 59–64.
6. Hampel, F.R., Ronchetti, E, Rousseeuw, P., Stahel, W. (1986). Robust statistics: The approach based on influence functions. Wiley, New York.
7. Handshin, E., Schweppe, F.C., Kohlas, J., Fiechter, A. (1975). Bad data analysis for power system state estimation. IEEE Transactions on Power Apparatus and Systems, Vol. PAS–94, 2, 329–337, with discussion.
8. Hill, R.W. and Holland, P.W. (1977). Two robust alternatives to least squares regression. JASA 72, 828–833.
9. Huber, P.J. (1981). Robust Statistics, Wiley, New York.
10. Kraśker, W.S. and Welsch, R.E. (1982). Efficient bounded influence regression estimation. JASA 77, 595–604.
11. Markatou, M. (1988). Robust bounded influence tests in linear models. Ph.D. Thesis, Penn. State University.

12. Maronna, R.A. and Yohai, V.T. (1981). Asymptotic behavior of general M–estimates for regression and scale with random carriers. Z. Wahrsch. verw. Beg. 58, 7–20.

13. Sen, P.K. (1982). On M–tests in linear models. Biometrika 69, 245–248.

14. Sheather, S.J. and Hettmansperger, T.P. (1987). Estimating the standard error of robust regression estimates. Proceedings of STATCOMP 1987, Melbourne, Australia, 26–39.

15. Welsch, R.E. (1980). Regression sensitivity analysis and bounded influence estimation. In: Evaluation of Econometric Models, J. Kmenta and J.B. Ramsey (eds.) Academic Press, New York, pp. 153–167.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 77	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Robust Bounded Influence Tests in Linear Models		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Marianthi Markatou, The University of Iowa Thomas P. Hettmansperger, The Pennsylvania State University		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C0741
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics The Pennsylvania State University University Park, PA 16802		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR042-446
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistical and Probability Program Code 436 Arlington, VA 22217		12. REPORT DATE November 1988 13. NUMBER OF PAGES 25
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Aligned generalized M-test, bounded influence, contiguous alternatives, influence function, linear model, random carriers, robustness, τ-test.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A robust test, which we call an aligned generalized, M-test for testing subhypotheses in the general linear models is developed, and its asymptotic properties are studied. The test is a robustification of the well known F-test and it is an elegant and practical alternative to Ronchetti's (1982) class of τ-tests. P-values associated with it can be approximated readily using existing chi square tables, unlike Ronchetti's test. The test is based on a appropriately constructed quadratic form, and uses the generalized M-estimators of the parameters in the reduced model. Under the null hypothesis the asymptotic

20. Abstract (Continued)

distribution is a central chi square, and under contiguous alternatives is a non-central chi square with the same degrees of freedom. The test can also be viewed as a generalization of Sen's (1982) M-test for linear models.

The influence function of the test is bounded. The bound not only applies to the influence of residuals but also to the influence of position in the factor space. Sen's test, on the other hand, has bounded influence only in residuals.